AFRL-RI-RS-TR-2017-237

DIVIDE AND RECOMBINE FOR LARGE COMPLEX DATA

STANFORD UNIVERSITY

DECEMBER 2017

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2017-237 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
RICHARD G. FEDORS
Work Unit Manager

/ S /
JULIE BRICHACEK
Chief, Information Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
DEC 2017	FINAL TECHNICAL REPORT	SEP 2012 – MAR 2017

4. TITLE AND SUBTITLE

DIVIDE AND RECOMBINE FOR LARGE COMPLEX DATA

5a. CONTRACT NUMBER
N/A

5b. GRANT NUMBER
FA8750-12-2-0343

5c. PROGRAM ELEMENT NUMBER
62702E

6. AUTHOR(S)
Patrick Hanrahan (Stanford University)
William Cleveland (Purdue University)
Jeffrey Heer (University of Washington)
Ryan Hafen (Pacific Northwest National Laboratory)

5d. PROJECT NUMBER
XD24

5e. TASK NUMBER

5f. WORK UNIT NUMBER
ROEM

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
1) The Board of Trustees of the Leland Stanford Junior University, 3160 Porter Drive, STE. 100 Palo Alto, CA 94304. 2) Purdue University, 250 N. University Street, Lafayette, IN 47907. 3) University of Washington, 1410 NE Campus Parkway, Seattle WA 98195

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RISC
525 Brooks Road
Rome NY 13441-4505

DARPA
675 North Randolph Street
Arlington VA 22203-2114

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2017-237

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Divide and Recombine (D&R) statistical approach was developed for analyzing 'big data' where the computational complexity is very high. The analyst divides data into subsets by a D&R division technique, applying analytic methods to each subset independently, without communication. Outputs of each analytic method are recombined by a D&R recombination procedure, which allows extensive parallel computation. DeltaRho software is the open-source implementation of D&R (see www.deltarho.org). Front end is the R package datadr, a language that makes programming D&R simple. At the back end running on a cluster, is a distributed database and parallel compute engine such as Hadoop, which spreads subsets and outputs across the cluster, and executes the analyst R and datadr code in parallel. The R package RHIPE provides communication between datadr and Hadoop. DeltaRho thus protects the analyst from having to manage the database and parallel computation. This research was performed under the XDATA program, to meet big data challenges by developing computational techniques and software tools for processing and analyzing vast amounts of mission-oriented information.

15. SUBJECT TERMS

Machine Learning, Visualization, Statistics, Big data, Parallel computing, Computational complexity, Hadoop, XDATA.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**RICHARD G. FEDORS**
U	U	U	UU	41	19b. TELEPHONE NUMBER (*Include area code*)

TABLE OF CONTENTS

SUMMARY

Data analyses, to be comprehensive and have low risk for missing critical information in the data, need to be deep. One property of deep analysis is that the data are, in part, analyzed in detail at their finest granularity. It is certainly true that analyzing summary statistics is also important. But analyzing just summary statistics, sometimes done to enable computation, present a high risk for Deep analysis for big data and high computational complexity of analytic methods. A new approach to data analysis is required that also provides for feasible, practical computation, which in turn, means analytic methods must be computed in parallel. Divide and Recombine (D&R) provides this. The DeltaRho software enables data analysts to carry out D&R in practice without having to manage the details of parallel computation, to that they can spend a large majority to their time thinking about the data and its analysis and a minority of their time programming the analysis. The analyst programs in R, the most used system for data analysis. Github is the development site for the DeltaRho software. The web site www.deltarho.org provides documentation for DeltaRho and how to download it.

The measure of success is gauged by analysis of big datasets with high computational complexity. We have ourselves had a number of successes. This has been demonstrated by our own analyses. Two examples are: 10,615,054,608 queries to the Spamhaus Internet Protocol (IP) address blocklisting service; and 50,632 3-hour satellite rain-rate measurements at 576,000 locations from the Tropical Rainfall Measurement Mission. We have run Computational Performance Measurement and Analysis (CPM&A) designed experiments that show fast performance. On a cluster with 10 nodes, 264 cores, 528 gigabytes (GB) of random access memory (RAM), and 88 terabytes (TB) of disk space we ran a logistic regression with 1 TB of data and 1 million subsets. We measured the elapsed time to read subsets into memory and form the subsets. We measured the elapsed time to carry out the computation of the logistic regressions of subsets and the recombination to get a single result. The two times were 12.1 minutes and 6.0 minutes, respectively, a total of 18.1 minutes. That is a practical amount of time for a data analyst to wait for a 1 TB dataset.

INTRODUCTION

In Divide and Recombine (D&R), the analyst divides the data into subsets by a D&R division method. Each analytic method is applied to each subset, independently, without communication. The outputs of each analytic method are recombined by a D&R recombination method. Sometimes the goal is one result for all of the data, such as a logistic regression; D&R theory and methods seek division and recombination methods to optimize the statistical accuracy. Much more common in practice is a division based on the subject matter. The data are divided by conditioning on variables important to the analysis. In this case the outputs can be the final result, or further analysis is carried out, an analytic recombination.

D&R computation is mostly embarrassingly parallel, the simplest parallel computation. DeltaRho software is an open-source implementation of D&R. (See www.deltarho.org.) The front end is the R package datadr, which is a language for D&R. It makes programming D&R simple. At the back end, running on a cluster, is a distributed database and parallel compute engine such as Hadoop, which spreads subsets and outputs across the cluster, and executes the analyst R and datadr code in parallel. The R package R and Hadoop Integrated Programming Environment (RHIPE) provides communication between datadr and Hadoop. DeltaRho protects the analyst from management of parallel computation and database management.

It is natural to divide data based on the subject matter. The data are divided by conditioning on variables important to the analysis. For example, in a collaboration between faculty and students in Purdue's Statistics Department, and Wen-wen Tung in the Earth, Atmospheric, Planetary Sciences Department and her students, we are analyzing a satellite dataset from the Tropical Rainfall Measuring Mission. There are 50,632 3-hour rainfall measurements at each of 576,000 locations. One division is by location across time, so there are 576,000 subsets with 50,632 measurements per subset. Another is by time across locations, so there are 50,632 subsets with 576,000 subsets per location. Subject matter division is just as valid for small datasets. It has been widely practiced in the past and is a statistical best practice. For D&R we use such division both for a best practice and for computational gain. Subject-matter division is the most used in practice.

In sampling division each subset is seen as a sample of the data. Subsets are replicate samples. For example, we can carry out random replicate division: choose subsets randomly. We seek a single result for all of the data. There is a statistical division method and a statistical recombination method. The statistical accuracy of the D&R result is typically less that that of the direct all-data result. D&R research in statistical theory seeks to maximize the statistical accuracy of D&R results. The accuracy depends on the division method and the recombination method. A community of researchers in this area is developing, not just those on our D&R team.

Computational performance of our software is a critical matter. However, today, Computational Performance Measurement and Analysis (CPM&A) for Big-Data and high computational complexity is often lacking in rigor, and not sufficiently informative. Performance tests typically use a few low-level computations such as sort, which are not informative for data analyses. We study response times of analytic methods, which are directly what an analyst uses and wants

to run as fast a possible. Benchmark testing today often fails to control for factors that are important for comparing aspects of two systems. For example, Hadoop configuration parameters can have a big impact. Pilot experiments show there are strong interactions among factors. Little attention today is given to interactions. We have been addressing these problems by multifactor experiments with many factors.

Experimentation is challenging. Knowledge of big-data systems is needed to determine salient factors. Interactions are not readily quantifiable from the knowledge. So empirical model building is necessary. Running a design on a cluster can be complex. Some factors have levels that can be changed only by system administrators. One current topic here is comparing the Spark back end and Hadoop; to keep factors fixed for each system, both must be installed on a cluster. Attention must be paid to limiting all other processes other than the operating system. Care must be taken to ensure replicate runs are independent and not affected by aspects of the systems that can create trends in duplicate runs. Many matters need to be evaluated by running on more that one cluster.

One critical aspect of our work is that through out the entire program we analyzed big datasets. Some of the datasets were those required by the XDATA program. But many others were those analyzed by members of the team who were matter experts in certain subject matter areas. These were live. "Live" means the success of the analysis is judged by the subject matter results. Live analyses serve as an effective heat engine for research ideas that solve real problems, and serve as a test bed for those ideas. Analyses need to be live because it is important to judge research based on how well it increases subject-matter knowledge and the effort required to get the increases. There are notions of statistical accuracy that can be important, too, but those become subject to the same criterion of increasing subject-matter knowledge.

Software For Divide and Recombine

Much of our work in XDATA focused on building software prototypes implementing the Divide & Recombine research. While in the XDATA program, we developed several iterations of components of a software stack for analysis and visualization of large complex data and used this to analyze the XDATA challenge datasets as well as our own research datasets. In addition to building the components, we also initiated a project, ultimately named DeltaRho, to encapsulate the entire body of work and build a community of users.

The components developed over the term of the XDATA program are the R and Hadoop Integrated Programming Environment (RHIPE), the datadr R package, the trelliscope R package, the next iteration trelliscopejs package, and an R interface to Anaconda, Inc.'s Bokeh visualization library, rbokeh. In the process of creating these packages and performing our XDATA data analysis work, several other related supporting software components were created. The effort for each of these components is described in more detail below.

datadr

The datadr R package provides a simple interface to D&R operations. It allows an analyst to operate wholly from within R on very large datasets, using R code to compute on the data. Instead of MapReduce, which is not a logical way to think for many data analysis tasks, datadr provides an interface to D&R, a simple concept very familiar to statisticians, while in the background mapping D&R tasks into the appropriate MapReduce steps. The datadr package provides generic interfaces to several D&R workflows, as well as constructs for reading writing data, computing common summary statistics, statistical distributions, and other aggregations in a divison-agnostic manner.

Our initial version of datadr required logic for plugging in different back ends such as Hadoop, local disk, or local memory, to be written from the ground up for each case, which made datadr less extensible and more difficult to maintain. In a second iteration, we created a back end agnostic interface, making it possible for datadr to harness new technology (such as Spark) much more straightforward as it comes along. Regardless of the back end, coding is done entirely in R and data is represented as R objects. Datadr currently supports in-memory, local disk / multicore, and Hadoop back ends.

We worked on experimental support for Apache Spark, demonstrating the back end agnostic design, but due to limitations in the R/Spark connector package, a solution ready for practical use was not reached.

RHIPE

RHIPE is the R and Hadoop Integrated Programming Environment. RHIPE allows an analyst to run Hadoop MapReduce jobs wholly from within R. RHIPE is used by datadr when the back end for datadr is Hadoop. You can also perform D&R operations directly through RHIPE, although in this case you are programming at a lower level. RHIPE existed before the XDATA program but throughout the course of XDATA it underwent many iterations and enhancements, including migrating to Hadoop 2.0, supporting several of the most popular Hadoop distributions, an improved build system, more and better documentation, and a large number of bug fixes. Work on fine-tuning Hadoop parameters with RHIPE was carried out through scaling tests, where we assessed the scalability of RHIPE by analyzing up to 128 terabytes of data on one of our institutional clusters.

trelliscope

Trelliscope is a D&R visualization tool based on Trellis Display that enables scalable, flexible, detailed visualization of data. Trelliscope, backed by datadr, scales Trellis Display, allowing the analyst to break potentially very large data sets into many subsets, apply a visualization method to each subset, and then interactively sample, sort, and filter the panels of the display on various quantities of interest. We built the trelliscope R package as a Shiny application that displays plots of divisions of a distributed data object, potentially backed by very large datasets sitting on Hadoop. In one of the summer challenges, we illustrated the scalability of this system by creating a Trelliscope display of NxCore data using over a million subsets representing over 50 billion data points.

trelliscopejs

The Trelliscope R package provides a web-based viewer written as a Shiny application. There are many limitations to this approach. One limitation is the need for a Shiny server to be able to deploy the visualizations to. Another major limitation is that Shiny is not well-suited for very complex UIs that require quick iterative interactivity within the browser. The application became unwieldy to maintain and very difficult to add new features to as the internals were so complicated. Because of these limitations, we rewrote Trelliscope from the ground up as a JavaScript viewing engine using the React framework. A lot of early experimentation was done with other JavaScript frameworks such as Ember and Angular. The trelliscopejs JavaScript library provides a pure-JavaScript web-based Trelliscope viewer which theoretically could be instantiated with data from any source. In our work, since we use R, we created an R interface to create and populate trelliscopejs displays. This R interface provides two approaches that fit seamlessly into a typical R users's workflow. One is the ability to easily create a Trelliscope display to an existing ggplot2 object. The other approach fits into the now-popular "tidyverse" ecosystem in R and allows users to build data frames of plots as "list-columns" of the data frame and easily generate a display from the data frame.

rbokeh

Toward the end of the XDATA program, we determined that the R world needed to have access to all the work going on in the Bokeh visualization library developed by Anaconda, Inc. We built an R interface to Bokeh, providing many of the features R users expect in a visualization system, drawing a lot of inspiration from ggplot2 in terms of providing a layered plot specification mechanism that provides features like mapping aesthetics, automatic legends, etc. This work allowed us to easily specify interactive graphics with the same ease as creating static plots with ggplot2. Much of the interactivity we get for free thanks to Bokeh, such as pan, zoom, and tooltips. More customized interactions were facilitated through callback support. We also built in support for rbokeh graphics to interact within Shiny applications.

graphqlr

While working on trelliscopejs, we anticipated a future where trelliscopejs would be able to more efficiently query the data it needs from R using Relay/GraphQL. GraphQL is a backend agnostic data query language and runtime that drastically reduces the number of server requests created by the browser by using a dynamic and nested query structure. To preprare for this, we developed the "gqlr" package. This package pulls inspiration from Facebook's graphql-js package and implements a full GraphQL server within R. "gqlr" allows R users to supply their own functions to satisfy the data requirements of a submitted GraphQL query, thus enjoying the rapid iteration time of R and production iteration time of GraphQL.

packagedocs

To help create a unified documentation website for all the packages in the DeltaRho stack, we developed the packagedocs R package, which automatically generates a web page of documentation and function reference pages based on R Markdown tutorial text and the R docs files. We used this package to build the documentation for the R package components of DeltaRho and served them with our website.

stlplus

The "stlplus" package was used extensively in many of the time series analysis we performed on XDATA challenge datasets as well as our own large research datasets. This package existed prior to XDATA as "stl2", but we migrated the code base to a revamped "stlplus" package. We ported the old C code to C++ using Rcpp as the integration mechanism. We revamped the documentation and got the package CRAN-ready and published it on the Comprehensive R Archive Network (CRAN).

rmote

To help deal with a common issue of working on a remote cluster head node from a local machine, but wanting to receive graphical output from that node while working within R, we built a package, "rmote". Traditional approaches to viewing graphics while working on remote machines includes X11 forwarding and using Virtual Network Computing (VNC) with a Linux desktop environment. Neither of these are very attractive solutions and have some major limitations.

We built the rmote to make working in R over secure socket shell (SSH) on a server a bit more pleasant in terms of viewing output. It opens a live connection from the remote machine to the local machine through websockets, opening a live-updating web page on the local machine that updates whenever a new output is produced from R. This setup allows a rich set of graphics to be passed back to the local machine, including web-based graphics, Shiny apps, etc.

devops

While our software work endeavored to make programming with big data as simple as possible, systems setup was always a major pain point. A great deal of development effort and support was focused on more simplified instantiation of Hadoop clusters. We created Vagrant images and dockerfiles for several different cluster scenarios supporting the DeltaRho stack. We also set up scripts to help easily provision a full-featured DeltaRho / Hadoop cluster on Amazon Web Services (AWS), and packaged the script up into a one-line setup tool.

Tessera / DeltaRho

A great deal of time and effort went into community building for our project. This began with the Tessera project, for which we created a website, tessera.io, which provided an introduction, installation and usage documents, the documentation for all the packages, and links to research. This project was superseded by the DeltaRho project, now hosted at deltarho.org.

Future / Impact

Our work on D&R research and software implementation has had a significant impact on the direction of statistical computing work for big data. Given the many talks and tutorials that we gave on the approach and the software over the course of the XDATA program, the ideas and software have been adopted by several researchers, and additional ideas and enhancements have been put forward by some of these groups. The datadr package provided the first distributed data object / distributed data frame representation in R which inspired further work in this area, including the "ddr" package developed out of a partnership that began at a big data workshop that we participated in at HP labs. Beyond this influence, our work on datadr and our general approach to big data has influenced work by RStudio on R/Spark. RStudio's approach is based on the "tidyverse" ecosystem, which early on was not compatible with the D&R approach.
A talk and discussions at the annual "Directions in Statistical Computing" workshop, which is organized by the R Core Team, led to the importance for RStudio to support arbitrary data structures and arbitrary R code execution in Spark, two necessary components for a D&R system to support. This work is still young, but is a promising future for big data in R and for D&R methods in R, with an open solution that is based on more modern technology (Spark) and supported by a more commercial and sustainable entity (RStudio). We hope that future D&R work can build upon this infrastructure. The trelliscopejs package has proven to be very useful independent of the rest of the DeltaRho ecosystem.

Scaling Interactive Visualization to Large Data Volumes

In the first phase of our XDATA award, we investigated methods for scaling interactive visualization techniques. Data analysts must make sense of increasingly large data sets, sometimes with billions or more records. We developed methods for interactive visualization of big data, following the principle that perceptual and interactive scalability should be limited by the chosen resolution of the visualized data, not the number of records. We mapped out a design space of scalable visual summaries that use data reduction methods (such as binned aggregation or sampling) to visualize a variety of data types. We then contributed methods for interactive querying (e.g., brushing & linking) among binned plots through a combination of multivariate data tiles and parallel query processing. We implemented our techniques in imMens, a browser-based visual analysis system that uses WebGL for both data processing and rendering on the graphics processing unit (GPU). In benchmarks imMens sustains 50 frames-per-second brushing & linking among dozens of visualizations, with invariant performance on data sizes ranging from thousands to billions of records. This work appeared at EuroVis 2013.

To support effective exploration, it is often stated that interactive visualizations should provide rapid response times. Indeed, this was the primary motivation for imMens. However, we found that the effects of interactive latency on the process and outcomes of exploratory visual analysis had not been systematically studied. In an experiment published at the Institute for Electrical and Electronic Engineers (IEEE) InfoVis 2014 symposium, we measured user behavior and knowledge discovery with interactive visualizations under varying latency conditions. We observed that an additional delay of 500 milli-seconds (ms) incurs significant costs, decreasing user activity and data set coverage. Analyzing verbal data from think-aloud protocols, we found that increased latency reduced the rate at which users made observations, drew generalizations and generated hypotheses. Moreover, we noted interaction effects in which initial exposure to higher latencies led to subsequently reduced performance in a low-latency setting. Overall, increased latency appears to cause users to shift exploration strategy, in turn affecting performance. Since publication this study has been highly cited, and used to help guide and motivate work on a number of scalable visualization and low-latency data processing systems.

Declarative Languages for Interactive Visualization: The Reactive Vega Stack

Another thread of XDATA research concerns systems and tools for interactive data visualization, particularly declarative languages for interactive visualization. Long a staple of database query languages such as Structured Query Language (SQL), and graphic design languages Hypertext Markup Language (HTML) and Cascading Style Sheets (CSS), declarative approaches to visualization have only more recently come to hold sway. Declarative models for visual encoding (mapping data to visual elements) have become a dominant way of expressing visualizations, providing the right balance of expressive power and usable, domain-specific constructs. While prior work supports declarative approaches to static visual encodings, custom interaction design is either unsupported or achieved only via imperative callbacks, starkly breaking with a declarative style. In this context, we contributed new models and corresponding system implementations that infuse declarative visual encoding approaches with abstractions for specifying interaction techniques.

Our first result was the development of a novel declarative model for interaction techniques. By treating user input as a streaming data source, we combined ideas from functional reactive programming and data stream processing to model interactive programs as continuous queries over data streams. We laid out this model in papers at the Association for Computing Machinery (ACM) User Interface and Software Technology (UIST) 2014 symposium, where we introduced our declarative model for interaction and demonstrated its expressivity, and IEEE InfoVis 2015, where we contributed an implementation of this model in the form of a reactive dataflow architecture. The Reactive Vega system supports optimized processing of streaming data and interactive updates, with superior performance to existing web-based visualization tools such as D3. The implementation involves a number of unique aspects, including the use of self-instantiating dataflows (a form of adaptive query plan) to handle data-dependent control flow.

While expressive and performant, the Reactive Vega language can still require verbose specifications: while control flow is handled by the model, the logic of event handling and data processing must be explicitly provided. As a result, Reactive Vega is not well-suited for quickly generating interactive plots in the midst of an analysis session. Considering this issue led us to a fundamental question. Given an assignment of a data field to a visual encoding channel (e.g., x-position, color, size), existing declarative visual encoding systems leverage a combination of data type information (e.g., quantitative, ordinal, or categorical) and smart defaults (e.g., perceptually effective color palettes) to automatically generate appropriate encoding logic. This facility lets users rapidly specify encodings, then override and modify default choices if desired. We asked: does an analogous relationship hold for interaction techniques? For example, given an indication of the semantics of the interaction, can one synthesize appropriate event handling logic?

This line of questioning led to the development of Vega-Lite, a concise high-level language for rapidly creating interactive, multi-view visualizations. We devised a new grammar of interactive graphics in which users specify interactions in terms of their semantics: the type of selection involved (e.g., point or interval selections) and associated transformations. By default, input event handling logic for populating these selections is automatically synthesized based on the selection type, applied transforms, and interaction context (e.g., mouse or touch-based), but remains customizable. This selection abstraction then "plugs in" to visual encodings to specify dynamic input data, drive conditional encoding logic, and define scale extents (e.g., for panning or zooming). Interactive selections in Vega-Lite enable an expressive set of interaction techniques in a surprisingly concise manner, with specifications one or more orders of magnitude more concise than prior work. In recognition of this work, we received the Best Paper Award at IEEE InfoVis 2016, a premier venue for Information Visualization research.

In terms of impact, beyond DARPA multiple companies, including Apple, Elastic Search, Fitbit, MapD and others, are using Vega open source software. Vega has been adopted as a standard for adding interactive visualizations to Wikipedia articles, and tools based on Vega-Lite have been developed in a variety of languages, most notably the Python Altair library for use within Jupyter Notebooks. While Vega and Vega-Lite are useful in their own right, our larger vision is to have these languages serve as a foundation for novel research. Towards this end, our group has built a number of research systems and models on top of the Vega stack:

- Lyra (EuroVis 2014) is a graphical interface for creating custom visualization designs without writing code, built on Vega and Vega-Lite. Lyra is more expressive than interactive systems like Tableau, allowing designers to create custom visualizations more comparable to hand-coded designs built with D3 or Processing. The resulting visualizations are realized as Vega specifications that can then be published and reused on the Web.

- Voyager (presented at IEEE InfoVis 2015 and ACM Conference on Human Factors in Computing Systems CHI 2017) is a visual analysis tool that combines both manual specification and automatic chart recommendation. Similar to other tools, Voyager users can specify charts manually by assigning data fields to visual encoding channels. However, in addition, Voyager suggests relevant visualizations based on a user's current focus. Underneath the hood, the CompassQL (Special Interest Group on Management of Data SIGMOD 2016 conference) visualization query language searches over thousands of potential Vega-Lite charts, ranks them using both statistical and perceptual measures, and presents the top-ranked examples in a recommendation gallery. In studies of early-stage data analysis, we found that this mode of exploration significantly increases data coverage compared to state-of-the-art tool designs. By leveraging Vega, any visualization in Voyager can be exported for further editing, including design customization in Lyra.

- GraphScape (ACM CHI 2017 Best Paper Nominee) provides a formal model of the relationships among Vega-Lite visualizations. GraphScape is a directed graph model, where nodes represent Vega-Lite charts and edges represent specification edits that turn one chart into another. GraphScape edges are weighted by an estimated "cost" of the effort needed to interpret one chart having seen another. This model enables applications such as automatic sequence generation for presentations and automatic search for design alternatives. For example, given an initial visualization design (such as a scatter plot), GraphScape can be used to find similar designs that better scale to large data volumes (e.g., binned aggregation or sampling). In controlled studies we have found that GraphScape sequence recommendations closely align with human judgments.o

Building DSLs for Data Analysis

The Stanford team worked on three main projects during the period of this grant. The first, Riposte, was a vector virtual machine for the R programming language. The second, Terra, was infrastructure for building high performance domain-specific languages. The third was Opt, A domain-specific language (DSL) for computing solutions to non-linear least squares problems. All these projects resulted in open sources releases and a body of dedicated users; all the systems were also documented with publications in premier conferences.

Riposte

Riposte is a virtual machines for array processing operations embedded in the R programming language. Vector virtual machines work well for long vectors. One of the most innovative features of the system is the optimizations for scalars and short vectors, which we term partial length specialization.

Another main focus of our research has been on adapting the code to the nature of the data. For example, if the data is grouped and aggregated; the method used depends on the number of groups and the number of elements in each group.

A significant amount of effort was also put into investigating new methods for building vector processing virtual machines. R implements primitive functions in C code as part of the interpreter itself, Riposte takes a more principled approach, implementing these functions in R code in an new standard library. Where necessary, Riposte defines some functions using C code through a new foreign function interface which permits specifying just the kernels of map, fold, and reduce-style operations in C. The Riposte interpreter handles the iteration over these kernels, allowing us to perform vector fusion over library-defined C code.

These new algorithms and approaches to building the system led to substantially faster implementations of many algorithms in R. We had hoped to integrate this work into the core R release, but the project was left partially unfinished because the team working on the project left for industry.

Terra

As part of this proposal we developed a new programming language Terra. Terra is a low-level C-like language embedded in lua. Lua is dynamically typed and has garbage collection. Terra is statically typed and memory management is explicit. The Terra sublanguage is based on LLVM (originally Low Level Virtual Machine, now an umbrella for compilation and machine code generation). We implement DSLs in Terra using meta-programming. Lua parses and analyzes the DSL, and then translates it into Terra. Terra is then jitted (just-in-time compilation) into very efficient low-level code.

One novel part of Terra is the concept of Exotypes, a type extension system. Exotypes create types that are as fast as native implementations in statically typed languages such as C, but are embedded in dynamically typed languages using a meta-object protocol. Using Exotypes, we have implemented Array(T), a type constructor that builds a high-performance array implementation where each element of the array is of type T. The Exotype compiler creates an efficient implementation using a combination of blocking and vectorization, much like ATLAS (automatically tuned linear algebra software) and other array processing meta-compilers use to achieve efficiency. Exotypes allow us to build systems like Riposte much more efficiently.

We use terralang.org for the documentation and portal into the system. Terra itself is maintained on github.

Opt

The final major project was a DSL called Opt designed to solve non-linear least squares optimization problems over large regular arrays or graph data structures. Users enter the function to be optimized as a simple expression representing an energy to be minimized. This expression has knowns and unknowns. The unknown variables are associated with different elements of the array or graph. Opt is implemented in Terra.

Opt automatically compiles an efficient solver for the given energy function. Our compiler automatically transforms these specifications into state-of-the-art GPU solvers based on Gauss-Newton or Levenberg-Marquardt methods. The compiler has several unique components. (1) We have designed a new method for optimizing symbolic tensor expressions based on Einstein summation notation. We have implemented a program analyzer for these expressions, and generate optimized code for different types of sparse matrices. (2) We have developed a method for auto-differentiation of symbolic tensor expressions.

We did an extensive evaluation of the system. This included comparing our approach to existing DSLs such as Ebb (developed at Stanford University) and Simit (developed at Massachusetts Institute of Technology). Opt solves for unknowns 100-600 times faster than previous approaches. The main reason Opt is that much faster is that it compiles a specialized solver that does not materialize a matrix (we call this a matrix-free solver).

Opt is available as open source and we have had roughly 100 downloads.

RESULTS AND DISCUSSION

WHAT DO WE GET FROM D&R? DEEP ANALYSIS

We get deep analysis, as described above, even when the data are big and computational complexity is high. This includes visualization of the detailed data, critical to statistical model building and validation, and to determining if a machine learning method is appropriate for the visualized patterns in the data. We do visualization by applying a method to subsets that have the detailed data. While it is feasible typically to apply a visualization method to all subsets, it is often not practical to look at them all because there can be far too many subsets, which in applications can be tens of thousands to the millions. So we sample. Sampling plans that preserve the information in the data can be readily devised because we can compute sampling variables across all subsets.

WHAT DO WE GET FROM D&R? HIGH COMPUTATIONAL PERFORMANCE

We get high computational performance. DeltaRho can increase dramatically the data size and analytic computational complexity that are feasible in practice, whether hardware power of an available cluster is small, medium, or large. The data can have a memory size that is larger than the physical cluster memory. For us this occurs routinely.

WHAT DO WE GET FROM D&R? ACCESS TO METHODS

We get access to the 1000s of methods of statistics, machine learning, and data visualization.

WHAT DO WE GET FROM D&R? HIGH EFFICIENCY PROGRAMMING BY THE ANALYST

We get very high efficiency in using R and datadr to program with the data, along with a great power and flexibility that allows deep analysis and tailoring analyses to the data. Most importantly, DeltaRho protects the analyst from the detail of distributed parallel computation and subset database management. Furthermore, datadr is abstracted from back end choices, so that its code is the same whatever the back end. For example, you can use datadr on a single multicore machine. Of course, back ends other than Hadoop require other software that connects datadr and a back end like RHIPE does for Hadoop.

WHAT DO WE GET FROM D&R? THE PRICE IS EXCELLENT

It's free. The software is all open source. See www.deltarho.org to get information on downloading the software, installing, and documentation to program datadr.

CONCLUSIONS

This project involved the development and evaluation of a new framework for data analysis, Divide and Recombine (D&R). In D&R, the analyst divides the data into subsets, applies an analytic method to each subset, and then combines the output of the analysis into a final result. The advantage of D&R is that it is easy to parallelize, and hence scalable to large datasets. D&R poses interesting theoretical problems in statistics, and this research project has developed ways to optimize the divide and recombine steps. Furthermore, an extensive suite of open source tools have been developed and released to the community. Finally, D&R has been used extensively in the XDATA project to perform data analysis of important datasets, and has shown to be a valuable tool in the data scientists toolchest.

The project also had two noteworthy, but smaller projects. The first involved new tools and methods for scalable visualization, and the second involved methods for developing domain specific languages for array processing and optimization. Again, all this software has been released to the community as open source.

APPENDIX

Publications	
Title Indicate the title of the publication.	imMens: Real-Time Visual Querying of Big Data
Author(s) Indicate the authors of the publication.	Zhicheng Liu, Biye Jiang and Jeffrey Heer
Publication Date Indicate the date of publication.	6/1/2013
Publication Venue Indicate the journal, conference, or magazine name.	Computer Graphics Forum (Proc. EuroVis'13)
Keywords Enter keywords for the publication.	scalable visualization, visual summary, big data
URL Enter the URL associated with this publication.	http://vis.stanford.edu/papers/immens

Title Indicate the title of the publication.	Riposte: A Trace-Drive Compiler and Parallel VM for Vector Code in R
Author(s) Indicate the authors of the publication.	Justin Talbot, Zachary DeVito, Pat Hanrahan
Publication Date Indicate the date of publication.	5/7/2013
Publication Venue Indicate the journal, conference, or magazine name.	Proceeding of the 21st International Conference on Parallel Architectures and Compilation Techniques (PACT'12)
Keywords Enter keywords for the publication.	

URL Enter the URL associated with this publication.	http://cs.stanford.edu/~zdevito/p43-talbot.pdf

Title Indicate the title of the publication.	Trelliscope: A System for Detailed Visualization in the Deep Analysis of Large Complex Data
Author(s) Indicate the authors of the publication.	Ryan Hafen, Luke Gosink, William S. Cleveland, Jason McDermott, Karin Rodland, and Kerstin Kleese-Van Dam
Publication Date Indicate the date of publication.	10/13/2013
Publication Venue Indicate the journal, conference, or magazine name.	IEEE Symposium on Large-Scale Data Analysis and Visualization (LDAV 2013)
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	http://ml.stat.purdue.edu/hafen/preprints/Hafen_LDAV_2013.pdf
Title Indicate the title of the publication.	The Design of an Efficient Vector Virtual Machine for Data Analytics
Author(s) Indicate the authors of the publication.	Justin Talbot
Publication Date Indicate the date of publication.	5/15/2013
Publication Venue Indicate the journal, conference, or magazine name.	PhD Thesis

Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	https://searchworks.stanford.edu/view/10159717

Title Indicate the title of the publication.	The Design of Terra: Harnessing the best features of high-level and low-level programming languages
Author(s) Indicate the authors of the publication.	Zach DeVito and Pat Hanrahan
Publication Date Indicate the date of publication.	5/1/2015
Publication Venue Indicate the journal, conference, or magazine name.	1st Summit oN Advances in Programming Languages, SNAPL 2015
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	snapl.org/2015/
Comments Enter any relevant comments about this publication.	This paper was selected for extended discussion at the inaugural summit on advances in programming languages.

Title Indicate the title of the publication.	Tessera: Open source software for accelerated data science
Author(s) Indicate the authors of the publication.	Sego LH, RP Hafen, HM Director, and RR LaMothe
Publication Date Indicate the date of publication.	4/1/2014

Publication Venue Indicate the journal, conference, or magazine name.	INMM Information Analysis Technologies, Techniques and Methods for Safeguards, Nonproliferation and Arms Control Verification Workshop
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	https://www.inmm.org/INMM/media/Documents/Presenations/Information%20Analysis/WorkshopProceedings2.pdf
Title Indicate the title of the publication.	Lyra: An Interactive Visualization Design Environment
Author(s) Indicate the authors of the publication.	Arvind Satyanarayan, Jeffrey Heer
Publication Date Indicate the date of publication.	4/1/2014
Publication Venue Indicate the journal, conference, or magazine name.	Computer Graphics Forum (Proc. EuroVis)
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	http://idl.cs.washington.edu/papers/lyra
Title Indicate the title of the publication.	Authoring Narrative Visualizations with Ellipsis
Author(s) Indicate the authors of the publication.	Arvind Satyanarayan, Jeffrey Heer

Publication Date Indicate the date of publication.	6/30/2014
Publication Venue Indicate the journal, conference, or magazine name.	Computer Graphics Forum (Proceedings EuroVis)
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	http://idl.cs.washington.edu/papers/ellipsis
Title Indicate the title of the publication.	First-class runtime generation of high-performance types using exotypes
Author(s) Indicate the authors of the publication.	Zachary DeVito, Daniel Ritchie, Matt Fisher, Alex Aiken, Pat Hanrahan
Publication Date Indicate the date of publication.	6/30/2014
Publication Venue Indicate the journal, conference, or magazine name.	PLDI '14 Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	http://dl.acm.org/citation.cfm?id=2594307
Title Indicate the title of the publication.	Just-in-time Length Specialization of Dynamic Vector Code
Author(s) Indicate the authors of the publication.	Justin Talbot, Zachary Devito and Pat Hanrahan

Publication Date Indicate the date of publication.	6/11/2014
Publication Venue Indicate the journal, conference, or magazine name.	Proceedings of ACM SIGPLAN International Workshop on Libraries, Languages, and Compilers for Array Programming (ARRAY'14)
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	http://cs.stanford.edu/~zdevito/riposte2.pdf

Title Indicate the title of the publication.	Declarative Interaction Design for Data Visualization
Author(s) Indicate the authors of the publication.	Arvind Satyanarayan, Kanit Wongsuphasawat, Jeffrey Heer
Publication Date Indicate the date of publication.	10/5/2014
Publication Venue Indicate the journal, conference, or magazine name.	ACM User Interface Software & Technology (UIST'14)
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2014-DeclarativeInteraction-UIST.pdf
Title Indicate the title of the publication.	Predictive Translation Memory: A Mixed-Initiative System for Human Language Translation

Author(s) Indicate the authors of the publication.	Spence Green, Jason Chuang, Jeffrey Heer, Christopher D. Manning
Publication Date Indicate the date of publication.	10/5/2014
Publication Venue Indicate the journal, conference, or magazine name.	ACM User Interface Software & Technology (UIST), October 2014
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2014-PTM-UIST.pdf
Title Indicate the title of the publication.	Human Effort and Machine Learnability in Computer Aided Translation
Author(s) Indicate the authors of the publication.	Spence Green, Sida Wang, Jason Chuang, Jeffrey Heer, Sebastian Schuster, Christopher D. Manning
Publication Date Indicate the date of publication.	11/1/2014
Publication Venue Indicate the journal, conference, or magazine name.	Empirical Methods in Natural Language Processing, October 2014
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2014-InteractiveTranslation-EMNLP.pdf

Title Indicate the title of the publication.	The Effects of Interactive Latency on Exploratory Visual Analysis
Author(s) Indicate the authors of the publication.	Zhicheng Liu, Jeffrey Heer
Publication Date Indicate the date of publication.	10/13/2014
Publication Venue Indicate the journal, conference, or magazine name.	IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2014-Latency-InfoVis.pdf
Title Indicate the title of the publication.	Learning Perceptual Kernels for Visualization Design
Author(s) Indicate the authors of the publication.	Cagatay Demiralp, Michael Bernstein, Jeffrey Heer
Publication Date Indicate the date of publication.	10/13/2014
Publication Venue Indicate the journal, conference, or magazine name.	IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	http://hci.stanford.edu/publications/2014/perceptualkernels/perceptualkernels-infovis2014.pdf

Title: Indicate the title of the publication.	Data Science: An Action Plan for the Field of Statistics
Author(s): Indicate the authors of the publication.	W.S. Cleveland
Publication Date: Indicate the date of publication.	11/27/2014
Publication Venue: Indicate the journal, conference, or magazine name.	Statistical Analysis and Data Mining
Keywords: Enter keywords for the publication.	
URL: Enter the URL associated with this publication.	http://onlinelibrary.wiley.com/
Comments: Enter any relevant comments about this publication.	Vol 7:414-417; reprinting of 2001 article in ISI Review, Vol 69

Title: Indicate the title of the publication.	Divide and Recombine (D&R): Data Science for Large Complex Data
Author(s): Indicate the authors of the publication.	W.S. Cleveland and R.P. Hafen
Publication Date: Indicate the date of publication.	12/31/2014
Publication Venue: Indicate the journal, conference, or magazine name.	Statistical Analysis and Data Mining
Keywords: Enter keywords for the publication.	
URL: Enter the URL associated with this publication.	http://ml.stat.purdue.edu/hafen/preprints/Cleveland_SADM_2014.pdf

Comments: Enter any relevant comments about this publication.	Vol 7:425-433
Title: Indicate the title of the publication.	Predictive Interaction for Data Transformation
Author(s): Indicate the authors of the publication.	Jeffrey Heer, Joseph Hellerstein, Sean Kandel
Publication Date: Indicate the date of publication.	1/04/2015
Publication Venue: Indicate the journal, conference, or magazine name.	7th Biennial Conference on Innovative Data Systems Research (CIDR'2015)
Keywords: Enter keywords for the publication.	
URL: Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2015-PredictiveInteraction-CIDR.pdf
Title: Indicate the title of the publication.	Forum77: An Analysis of an Online Health Forum Dedicated to Addiction Recovery
Author(s): Indicate the authors of the publication.	Diana MacLean, Sonal Gupta, Anna Lembke, Christopher D. Manning, Jeffrey Heer
Publication Date: Indicate the date of publication.	12/31/2014
Publication Venue: Indicate the journal, conference, or magazine name.	ACM Computer-Supported Cooperative Work (CSCW), 2015

Keywords: Enter keywords for the publication.	
URL: Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2015-Forum77-CSCW.pdf
Comments: Enter any relevant comments about this publication.	* Best Paper Honorable Mention *

Title: Indicate the title of the publication.	Perfopticon: Visual Query Analysis for Distributed Databases
Author(s): Indicate the authors of the publication.	Dominik Moritz, Daniel Halperin, Bill Howe, Jeffrey Heer
Publication Date: Indicate the date of publication.	3/31/2015
Publication Venue: Indicate the journal, conference, or magazine name.	Computer Graphics Forum (Proc. EuroVis), 34(3), 2015
Keywords: Enter keywords for the publication.	
URL: Enter the URL associated with this publication.	http://idl.cs.washington.edu/papers/perfopticon
Title: Indicate the title of the publication.	Refinery: Visual Exploration of Large, Heterogeneous Networks through Associative Browsing
Author(s): Indicate the authors of the publication.	Sanjay Kairam, Nathalie Henry Riche, Steven Drucker, Roland Fernandez, Jeffrey Heer

Publication Date: Indicate the date of publication.	3/31/2015
Publication Venue: Indicate the journal, conference, or magazine name.	Computer Graphics Forum, Eurographics Conference on Visualization (Proc. EuroVis), 34(3), 2015
Keywords: Enter keywords for the publication.	
URL: Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2015-Refinery-EuroVis.pdf

Title Indicate the title of the publication.	A Demonstration of the BigDAWG Polystore System
Author(s) Indicate the authors of the publication.	Aaron Elmore, Jennie Duggan, Michael Stonebraker, Magdalena Balazinska, Ugur Cetintemel, Vijay Gadepally, Jeffrey Heer, Bill Howe, Jeremy Kepner, Tim Kraska, Samuel Madden, David Maier, Timothy Mattson, Stavros Papadopoulos, Jeff Parkhurst, Nesime Tatbul
Publication Date Indicate the date of publication.	8/12/2015
Publication Venue Indicate the journal, conference, or magazine name.	Proc. Very Large Database Endowment (PVLDB), 8(12), 2015
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	http://idl.cs.washington.edu/papers/bigdawg-demo

Title Indicate the title of the publication.	Natural Language Translation at the Intersection of AI and HCI

Author(s) Indicate the authors of the publication.	Spence Green, Jeffrey Heer, Christopher D. Manning
Publication Date Indicate the date of publication.	9/1/2015
Publication Venue Indicate the journal, conference, or magazine name.	Communications of the ACM, 58(9), pp. 46-53, 2015
Keywords Enter keywords for the publication.	
URL Enter the URL associated with this publication.	http://idl.cs.washington.edu/papers/translation-ai-hci

Title Indicate the title of the publication.	Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations
Author(s) Indicate the authors of the publication.	Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, Jeffrey Heer
Publication Date Indicate the date of publication.	1/15/2016
Publication Venue Indicate the journal, conference, or magazine name.	IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis'15), Jan 2016
Keywords Enter keywords for the publication.	Data voyager, visualization tools, rapid exploration
URL Enter the URL associated with this publication.	http://idl.cs.washington.edu/papers/voyager

Title Indicate the title of the publication.	Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation
Author(s) Indicate the authors of the publication.	Matthew Kay, Jeffrey Heer
Publication Date Indicate the date of publication.	1/25/2016
Publication Venue Indicate the journal, conference, or magazine name.	IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis'15), Jan 2016
Keywords Enter keywords for the publication.	perceptual "laws", visualizations, correlations
URL Enter the URL associated with this publication.	http://idl.cs.washington.edu/papers/beyond-webers-law

Title Indicate the title of the publication.	Reactive Vega: A Streaming Dataflow Architecture for Declarative Interactive Visualization
Author(s) Indicate the authors of the publication.	Arvind Satyanarayan, Ryan Russell, Jane Hoffswell, Jeffrey Heer
Publication Date Indicate the date of publication.	1/1/2016
Publication Venue Indicate the journal, conference, or magazine name.	IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis'15), Jan 2016
Keywords Enter keywords for the publication.	Reactive Vega, streaming, data
URL Enter the URL associated with this publication.	http://idl.cs.washington.edu/papers/reactive-vega-architecture

Title Indicate the title of the publication.	The VERP Explorer: A Tool for Exploring Eye Movements of Visual-Cognitive Tasks Using Recurrence Plots
Author(s) Indicate the authors of the publication.	Cagatay Demiralp, Jesse Cirimele, Jeffrey Heer, Stuart K. Card
Publication Date Indicate the date of publication.	10/15/2015
Publication Venue Indicate the journal, conference, or magazine name.	Workshop on Eye Tracking and Visualization (ETVIS), Oct 2015
Keywords Enter keywords for the publication.	eye movements, visual-cognitive tasks, VERP Explorer
URL Enter the URL associated with this publication.	http://idl.cs.washington.edu/papers/verp

Title Indicate the title of the publication.	Visual Debugging Techniques for Reactive Data Visualization
Author(s) Indicate the authors of the publication.	Jane Hoffswell, Arvind Satyanarayan, Jeffrey Heer
Publication Date Indicate the date of publication.	6/6/2016
Publication Venue Indicate the journal, conference, or magazine name.	Computer Graphics Forum, EuroVis 2016
Keywords Enter keywords for the publication.	

URL Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2016-VegaDebugging-EuroVis.pdf
Title Indicate the title of the publication.	Opt: A Domain Specific Language for Non-linear Least Squares Optimization in Graphics and Imaging
Author(s) Indicate the authors of the publication.	Zachary DeVito, Michael Mara, Michael Zollhofer, Gilbert Bernstein, Jonathan Ragan-Kelley, Christian Theobalt, Pat Hanrahan, Matthew Fisher, Matthias Niessner
Publication Date Indicate the date of publication.	4/10/2016
Publication Venue Indicate the journal, conference, or magazine name.	arxiv.org Subsequenctly published in ACM Transactions on Computer Graphics (Volume 36 Issue 5, October 2017)
Keywords Enter keywords for the publication.	optimization, non-linear least-squares, DSL
URL Enter the URL associated with this publication.	http://arxiv.org/abs/1604.06525 https://dl.acm.org/citation.cfm?id=3132188
Title Indicate the title of the publication.	Divide and Recombine: Approach for Detailed Analysis and Visualization of Large Complex Data
Author(s) Indicate the authors of the publication.	Ryan Hafen
Publication Date Indicate the date of publication.	2/18/2016
Publication Venue Indicate the journal, conference, or magazine name.	Book Chapter (Handbook of Big Data)

Keywords Enter keywords for the publication.	big data, exploratory, visualization
URL Enter the URL associated with this publication.	

Title Indicate the title of the publication.	Towards A General-Purpose Query Language for Visualization Recommendation
Author(s) Indicate the authors of the publication.	Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe and Jeffrey Heer
Publication Date Indicate the date of publication.	6/1/2016
Publication Venue Indicate the journal, conference, or magazine name.	ACM SIGMOD Human-in-the-Loop Data Analysis Workshop, 2016
Keywords Enter keywords for the publication.	Human-in-the-Loop Data Analysis
URL Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2016-CompassQL-HILDA.pdf
Comments Enter any relevant comments about this publication.	ACM SIGMOD Human-in-the-Loop Data Analysis (HILDA) Workshop, 2016
Title Indicate the title of the publication.	Surprise! Bayesian Weighting for De-Biasing Thematic Maps
Author(s) Indicate the authors of the publication.	Michael Correll, Jeffrey Heer.
Publication Date Indicate the date of publication.	7/1/2016

Publication Venue Indicate the journal, conference, or magazine name.	Accepted to IEEE InfoVis 2016
Keywords Enter keywords for the publication.	De-Biasing Thematic Maps
URL Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2017-SurpriseMaps-InfoVis.pdf
Title Indicate the title of the publication.	Vega-Lite: A Grammar of Interactive Graphics
Author(s) Indicate the authors of the publication.	Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, Jeffrey Heer.
Publication Date Indicate the date of publication.	7/01/2016
Publication Venue Indicate the journal, conference, or magazine name.	Accepted to IEEE InfoVis 2016 Subsequently published in IEEE Transactions on Visualization and Computer Graphics (Volume 23 Issue 1, January 2017)
Keywords Enter keywords for the publication.	Vega-Lite Interactive Graphics
URL Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2017-VegaLite-InfoVis.pdf

Title Indicate the title of the publication.	GraphScape: A Model for Automated Reasoning about Visualization Similarity and Sequencing
Author(s) Indicate the authors of the publication.	Younghoon Kim, Kanit Wongsuphasawat, Jessica Hullman, Jeffrey Heer

Publication Date Indicate the date of publication.	1/5/2017
Publication Venue Indicate the journal, conference, or magazine name.	ACM Human Factors in Computing Systems (CHI)
Keywords Enter keywords for the publication.	GraphScape Sequencing Visualization
URL Enter the URL associated with this publication.	http://idl.cs.washington.edu/files/2017-GraphScape-CHI.pdf
Title Indicate the title of the publication.	Voyager 2: Augmenting Visual Analysis with Partial View Specifications
Author(s) Indicate the authors of the publication.	Kanit Wongsuphasawat, Zening Qu, Dominik Moritz, Riley Chang, Felix Ouk, Anushka Anand, Jock Mackinlay, Bill Howe, Jeffrey Heer
Publication Date Indicate the date of publication.	5/06/2017
Publication Venue Indicate the journal, conference, or magazine name.	ACM Human Factors in Computing Systems (CHI)
Keywords Enter keywords for the publication.	Augment Computing Analysis
URL Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2017-Voyager2-CHI.pdf

Title Indicate the title of the publication.	Value-Suppressing Uncertainty Maps

Author(s) Indicate the authors of the publication.	Michael Correll, Dominik Moritz, Jeffrey Heer
Publication Date Indicate the date of publication.	9/4/2017
Publication Venue Indicate the journal, conference, or magazine name.	In submission to IEEE InfoVis 2017
Keywords Enter keywords for the publication.	Uncertainty Maps Visualization
URL Enter the URL associated with this publication.	
Title Indicate the title of the publication.	Extracting and Retargeting Color Mappings from Bitmap Images of Visualizations
Author(s) Indicate the authors of the publication.	Jorge Poco, Angela Mayhua, Jeffrey Heer
Publication Date Indicate the date of publication.	8/29/2017
Publication Venue Indicate the journal, conference, or magazine name.	Submitted to IEEE InfoVis 2017. E-published in IEEE Transactions on Visualization and Computer Graphics.
Keywords Enter keywords for the publication.	Bitmap Mappings Visualizations
URL Enter the URL associated with this publication.	https://www.ncbi.nlm.nih.gov/pubmed/28866538
Title Indicate the title of the publication.	Augmenting Code with In Situ Visualizations to Aid Program Understanding

Author(s) Indicate the authors of the publication.	Jane Hoffswell, Arvind Satyanarayan, Jeffrey Heer
Publication Date Indicate the date of publication.	10/22/2017
Publication Venue Indicate the journal, conference, or magazine name.	In submission to ACM UIST 2017
Keywords Enter keywords for the publication.	Augmenting Visualizations In Situ
URL Enter the URL associated with this publication.	
Title Indicate the title of the publication.	GraphScape: A Model for Automated Reasoning about Visualization Similarity and Sequencing
Author(s) Indicate the authors of the publication.	Younghoon Kim, Kanit Wongsuphasawat, Jessica Hullman, Jeffrey Heer
Publication Date Indicate the date of publication.	5/6/2017
Publication Venue Indicate the journal, conference, or magazine name.	ACM Human Factors in Computing Systems (CHI), 2017
Keywords Enter keywords for the publication.	Visualization Sequencing Automation
URL Enter the URL associated with this publication.	
Comments Enter any relevant comments about this publication.	Best Paper Honorable Mention Award

Title Indicate the title of the publication.	Regression by Eye: Estimating Trends in Bivariate Visualizations
Author(s) Indicate the authors of the publication.	Michael Correll, Jeffrey Heer
Publication Date Indicate the date of publication.	5/6/2017
Publication Venue Indicate the journal, conference, or magazine name.	ACM Human Factors in Computing Systems (CHI), 2017
Keywords Enter keywords for the publication.	Regression Bivariate Visualizations
URL Enter the URL associated with this publication.	
Title Indicate the title of the publication.	Reverse-Engineering Visualizations: Recovering Visual Encodings from Chart Images
Author(s) Indicate the authors of the publication.	Jorge Poco, Jeffrey Heer
Publication Date Indicate the date of publication.	6/12/2017
Publication Venue Indicate the journal, conference, or magazine name.	EuroVis 2017 Conference
Keywords Enter keywords for the publication.	Images Encoding Reverse-Engineering
URL Enter the URL associated with this publication.	https://idl.cs.washington.edu/files/2017-ReverseEngineeringVis-EuroVis.pdf